DOCUMENT RESUME

ED 209 247 TM 810 635

TITLE	Test Use and Validity: A Response to Charges in the Nader/Nairn Report on ETS.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	Feb 80
NOTE	27p.; For a related document, see TM 810 634.
AVAILABLE FROM	Educational Testing Service, Publications Order Services, Dept. E01, Princeton, NJ 08541 (Free)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Admission Criteria; *College Admission; *College Entrance Examinations; Higher Education; Predictive Validity; Scores; *Test Use; *Test Validity
IDENTIFIERS	Educational Testing Service; *Reign of ETS (Nairn); *Scholastic Aptitude Tes

ABSTRACT

The Nairn report, The Reign of ETS, has charged that the major college admissions tests administered by Educational Testing Service (ETS) have undue influence on admissions to higher education, and that the tests have little value in predicting future academic performance. Nairn's claims that the Scholastic Aptitude Test (SAT) is a poor predictor of performance in college and are based on faulty statistics. He uses an incorrect value for the characteristic validity of the SAT (.345) since he mistakenly averages the separate validities of the two parts of the SAT, rather than considering the validity of the whole test (.41). Predictions based on valid information will be better than random predictions. The best predictor of college grades is the high school record, but the SAT is nearly as good, and the two together are better than either alone. The tests administered by ETS are sponsored and controlled by associations of colleges that use scores on the tests as one factor in admissions decision-making. They are designed to provide a common basis for evaluation, to supplement students' academic records, and to permit students to satisfy admissions testing requirements through taking a single examination. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED209247

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H.C. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TEST USE AND VALIDITY

A Response to Charges in the Nader/Nairn Report on ETS

TM 810 625

Educational Testing Service • Princeton, New Jersey • February 1980

AN OVERVIEW

TEST USE AND VALIDITY

A recent report on ETS, written by Allan Nairn and sponsored by Ralph Nader, has charged that the major admissions tests administered by ETS have undue influence on admissions to higher education, and that the tests have little value in predicting future academic performance. The Nairn report arrives at its remarkable conclusions by misrepresenting the purposes and uses of the tests, and by distorting or ignoring the results of research.

Nairn's claims that the SAT is a poor predictor of performance in college are based on faulty statistics. He uses an incorrect value for the characteristic validity of the SAT (.345) since he mistakenly averages the separate validities of the two parts of the SAT, rather than considering the validity of the whole test (.41). After squaring that coefficient and doing some further arithmetic, Nairn comes to the conclusion that "for 88% of the applicants . . . an SAT score will predict their grade rank no more accurately than a pair of dice."

Nairn's interpretation of this statistic is clearly wrong and would make no more sense to a layman than to a professional statistician. Even if the SAT had no predictive validity, it would be *no worse* than a random predictor. However, predictions based on valid information will be better than random predictions, and will certainly be more accurate more than 50 percent of the time. Yet Nairn claims that dice will be as good a predictor as the SAT for 88 percent of the applicants. Nairn's dice should be carefully inspected.

Admissions test scores are used with students' previous grades and other information in predicting later academic performance. The best predictor of college grades is the high school record, but the SAT is nearly as good, and the two together are better than either alone. At the graduate and professional school levels, admission test scores are most often the best available predictors of later grades.

Copyright © 1980 by Educational Testing Service. All rights reserved.

Nairn charges that exaggerated and untrue claims are made for the tests. In support of this charge, Nairn quotes incorrectly from ETS publications. In one instance, he creates a quote, which he attributes to ETS, from parts of two separate and conflicting statements about aptitude testing from two different sources. They were used by ETS to illustrate the wide difference of opinion on that subject. Nairn also quotes from an SAT publication, deleting words and adding others in brackets to transform the meaning of the original passage. Nairn concludes that testers have assumed a right "to define the potential for thinking." The fact is that ETS claims only that the tests can be a useful measure of knowledge or academic ability. "No test," state College Board bulletins and manuals, "predicts with any certainty 'success in life' or is in any way a measure of an individual's total worth."

The tests administered by ETS are sponsored and controlled by associations of colleges and graduate and professional schools that use scores on the tests as one factor in admissions decision-making. The tests that are described as aptitude tests were never meant to assess innate, unchanging abilities, but to measure learned skills. They are designed to provide a common basis for evaluation, to supplement students' academic records, and to permit students to satisfy admissions testing requirements of institutions across the country through taking a single examination at nearby testing centers, at relatively low cost.

No doubt improvements in admissions can be made. But Nairn does not suggest alternatives that would perform the important functions now performed by admissions tests. Because of this, and because of its misrepresentations, the Nairn report is unlikely to contribute to the improvement of admissions to higher education.

TEST USE AND VALIDITY

AN INTRODUCTORY NOTE

The report on ETS commissioned by Ralph Nader and prepared by Allan Nairn sets out to discredit the usefulness of tests by attempting to prove, among other negative "findings," that test scores bear little relation to success in college or university. This conclusion will be puzzling to many people since it is contrary to the practical experience of educators and to the hundreds of research studies made all over the country through several decades.

The issue is of importance to educators and especially to the students who need to know whether or not the scores are in fact indicative of probable college success. ETS has always said—as have others with experience in this area—that the scores are by no means infallible guides and should always be used as only one factor in any judgment about whether a student is likely to do well in college or graduate studies. For many years up until the Nairn analysis, ETS and others have interpreted the evidence as showing that the student's previous grades are the most important indicators, but that the addition of scores on good tests adds significantly to the prediction of success.

This paper presents a careful review of the Nairn analysis and finds that in fact it is incorrect: as experience and research have told us for some time, test scores do indeed contribute useful information on which to base admissions decisions. The Nairn exercise manages to arrive at the opposite conclusions through a combination of selective reporting and statistical error, and presents them with a rhetorical flourish that fits the evident intent to demolish the tests rather than enlighten the reader.

It is a matter of public concern that educational decisions be based on the most valid available information. Clearly, advocacy in the areas of testing and admission is higly desirable as long as it is based on a fair presentation of fact. It would be a disservice to students, parents, institutions, and the public at large, however, to leave unanswered an attack that marshalls half truths to create an advocate's case against the use of test scores as one element in admissions. This paper is intended to help clarify some of the important issues that may have been confused by the Nader/Nairn report.

William W. Turnbull
President
Educational Testing Service

TEST USE AND VALIDITY

A recent report on ETS, written by Allan Nairn and sponsored by Ralph Nader, has charged that the major admissions tests administered by ETS on behalf of associations of colleges and graduate and professional schools constitute a "respectable fraud." The report asserts that exaggerated claims have been made for the tests, that as a result the tests have undue influence on admissions to higher education, and that in fact the tests have little value in predicting students' future academic performance. According to Nairn, most of the time admissions officers could do as well by rolling dice.

This charge will come as a surprise to the colleges and universities that have used scores on these tests as one element in admissions for decades. The faculty and admissions officers at these institutions know what the tests are designed to do and how they are used. Moreover, they have vast experience in selecting students for admission, in teaching these students, and in observing their academic performance. Apart from this experience, colleges and universities have conducted well over one thousand formal research studies of the relationship of test scores to the grades later obtained by enrolled students. Many of these studies have been conducted in collaboration with ETS, others by the institutions independently. *These studies show clearly that scores on the tests are useful in predicting the grades that students will obtain.*

The Nairn report arrives at its remarkable conclusions by misrepresenting the purposes and uses of the tests and by distorting or ignoring the results of research. Despite the fact that accurate information on these matters has been widely published and available to test users and to the public for a long time, the attention that the Nairn report will attract through its exaggerated and sensational charges requires that the record be set straight.

What the Tests Are Designed to Do

The major admissions tests administered by ETS are sponsored and controlled by associations of colleges and graduate and professional schools that use scores on the tests as one factor in admissions decision-making. Because grades and other information used in admissions (such as recommendations) are not directly comparable from student to student, these institutions ask that applicants also submit test scores, which supplement the other information and provide a more uniform basis for evalution of students' academic abilities and achievement. The institutions have long recognized the advantages to themselves in agreeing to accept results from uniform

systems of professionally developed examinations and the advantages to their applicants in having results from a single examination, administered nationwide, acceptable for admissions purposes at multiple institutions. In fact, recognition of the benefits to applicants came first; the College Board was founded in 1900 by a group of colleges in response to the problems of secondary schools and their students in meeting different and uncoordinated college admission testing requirements.

The tests themselves vary in their characteristics. The major testing program at the college level, the College Board Admissions Testing Program, offers a test of scholastic aptitude (the SAT), a test of English usage designed primarily for decision-making about course placement of entering students, and a series of subject matter Achievement Tests. The Graduate Record Examinations (GRE) program similarly offers an aptitude test and achievement tests in areas of graduate study. Law schools and graduate management schools generally use tests that may be broadly characterized as aptitude tests—the Law School Admission Test (LSAT) and the Graduate Management Admission Test (GMAT).

Since the tests that are described as aptitude tests are most widely used, most attention focuses on them. A common misconception is that these tests somehow measure innate, unchanging abilities. In fact, they measure learned skills. They are described as aptitude tests because they are not tied to a particular course of study, curriculum or program, and because they are typically used to assess students' relative abilities to perform well in future academic work. Scholastic aptitude tests are most often made up of problems that test reading comprehension, verbal reasoning and vocabulary, mathematical reasoning, and data interpretation. They measure intellectual skills that students are expected to have developed through both school and non-school experiences, apart from the particular courses of study they may have pursued. These skills are also rightfully regarded as broadly applicable to success in a variety of future courses of study, since higher education at all levels requires ability to read with understanding, to reason with words and numbers, and to calculate and understand quantitative relations.

The philosophy underlying much of education in the United States today places great stress on providing students with a broad base of general education and on delay of academic or professional specialization until a relatively late age. Students throughout high school and college are encouraged to acquire a broad acquaintance with the liberal arts and science. Considerable emphasis is placed on exploration, flexibility, and choice. There is no national syllabus. In this context, using tests of developed abilities, rather than relying exclusively on subject matter achievement tests at points of

transition from school to college or from college to graduate school, is a consistent, appropriate, and sound policy.

Each of the major admissions tests is made up of multiple-choice questions. There are principally three reasons for this. *First*, use of these questions permits a broad sampling of problems of different kinds in limited testing time. Other forms of assessment cannot cover as much material in any given time, greatly increasing the risk that students will be evaluated in terms of only a few topics that do not fairly represent what they know. *Second*, many studies have shown that evaluation of essays or other free response exercises is fraught with unreliability. In practice, each paper can be read by only a few individuals, and the grades assigned by different readers to any single paper tend to vary a great deal. This also poses risks to students, since their scores will be affected by the standards and points of view of the particular readers to whom their papers happen to be assigned. In contrast, the questions on well-developed, multiple choice tests are reviewed by many individuals, and there is a consensus on the correct answers; scores on tests consisting of a number of such questions tend to be much more reliable than scores on essay tests. *Third*, multiple-choice testing and the automated scoring it permits have a number of practical advantages. The low costs of the major testing programs to students and timely reporting of information are, in large part, a result of this form of testing.

In short, the major admissions tests are designed to provide a common basis for evaluation, supplementary to students' academic records which vary in meaning from school to school; a reliable and fair means of assessing relative student abilities to perform well in future academic work, one which is not heavily dependent on the specific nature of students' previous academic preparation; and a means for students to satisfy admissions testing requirements of institutions across the country through taking a single examination at nearby testing centers and at relatively low cost.

Claims for the Tests

Nairn's charge of fraud is largely based on a premise that exaggerated and untrue claims are made for the tests. For example, Nairn asserts:

> When ETS' claims about its aptitude tests are juxtaposed with evidence from scholarly studies and internal documents, it becomes difficult to conclude that they constitute an even minimally prudent construction of the facts. (p. 57)
>
> The ETS consumer is a victim of a false representation of a product . . . This respectable fraud is based on a benignly confident adherence to a system of psychometric belief. It is a

> system characterized by a tolerance for obfuscation and a bedrock assumption about the right and responsibility of the mental tester to use multiple-choice questions to define the potential for thinking, and then, regardless of the evidence, to draw broad conclusions about the extent to which a person possesses it. (p. 58)

In support of this charge, Nairn draws on incomplete quotations taken from ETS publications. In some cases these quotations are not only taken out of context and modified, but are incorrectly attributed to ETS. For example, the section of Nairn's report from which these statements are quoted begins with the following sentences:

> According to an ETS technical manual, "an aptitude test is . . . a device for measuring the capacity or potentiality of an individual for a particular kind of behavior." Such a device, says ETS, is "used to predict success in some occupation or training course." (p. 55)

The words in the original text that have been removed and reduced to an ellipsis (. . .) are: "commonly thought of as." The context from which these statements were taken is as follows:

> The field of measurement, of course, does not offer any single, universally accepted definition of an aptitude test. Of several which have been offered, perhaps the sharpest contrast is afforded by the following. According to Cronbach (1960, p. 20), "An *aptitude test* is one used to predict success in some occupation or training course . . . In form, these tests are not distinctly different from other types . . . The test is referred to as an achievement test when it is used primarily to examine a person's success in past study, and as an aptitude test when it is used to forecast his success in some future course or assignment." However, Ryans and Frederiksen (1951, p.456) find that "An *aptitude test* is commonly thought of as a device for measuring the capacity or potentiality of an individual for a particular kind of behavior. In the measurement of aptitude, previous experiences or training on the part of the individual is assumed either to be lacking or to be constant for all individuals comprising the population considered."
>
> The Cronbach view, of course, is extremely functional. While the SAT obviously meets this functional criterion, its development over the years since 1926 has also been guided largely by a conception similar to the Ryans and Frederiksen definition, that "previous experience or training . . . is assumed . . . to be constant for all individuals." (Angoff, 1971, p. 16)

Other statements attributed to ETS and quoted in this section of the report are contained in the following text by Nairn:

> The SAT student bulletin informs teenagers that the test can "measure [their] ability to understand what [they] read . . . the extent of [their] vocabulary . . . [and other] abilities [which] have been shown to be related to college work." The LSAT student bulletin asserts that the test is a reliable measure "of certain mental abilities related to academic performance in law schools." In 1977 the GRE student bulletin announced that the GRE Aptitude Test had been reissued in a new, improved model which would "measure analytical ability as well as . . . verbal and quantitative ability." (p. 56)

For the record, the full description of the SAT excerpted by Nairn was as follows:

> The SAT is a multiple-choice test made up of separately timed verbal and mathematical sections. Verbal questions measure your ability to understand what you read and the extent of your vocabulary. Mathematical questions measure your ability to solve problems involving arithmetic reasoning, algebra, and geometry. These abilities have been shown to be related to successful academic performance in college. *The SAT does not measure other kinds of abilities which may be associated with success in college*, such as special talents or motivation.
>
> Your high school record is probably the best evidence of your preparation for college. Because applicants have taken different courses and come from high schools with different grading practices, college admissions officers need a common measure of ability, such as the SAT. However, scores on the SAT are just part of the information used in making an admission decision. (*Taking the SAT*, 1978, p. 3, emphasis added)

These are but a few examples of the lack of care and forthrightness by Nairn in the use of quotations from published sources. Other statements (which Nairn quotes selectively later in his report) may be found in guidelines published in College Board bulletins and manuals for students, high school counselors, and admissions officers. For example, these guidelines include the following statements:

- test scores, like all types of measurements, physical as well as psychological, are not perfectly precise and should not be treated as though they were;
- although admissions test scores are good predictors of performance in college, they are not infallible predictors;
- tests can be a useful measure of knowledge or academic ability, but no test predicts with any certainty "success in life" or is in

any way a measure of an individual's total worth.

In the same guidelines, admissions officers are advised:

> Test scores should not be the sole factor in determining the admission of an applicant to an institution but should be considered one aspect of the description of an individual in this process. Data from the College Board's Admissions Testing Program is intended to supplement the secondary school record and other relevant information about the student in assessing competence for college work . . . Colleges should view admissions tests scores as approximate indicators rather than exact measures of a student's abilities and/or achievement. (*Taking the SAT*, 1978, back cover; *ATP Guide for High Schools and Colleges*, 1979, inside back cover.)

These statements, even after selection and editorial surgery by Nairn, clearly indicate that the tests measure particular, defined skills that have a bearing on future academic performance and that inferences about these skills cannot be made with any certainty. How Nairn arrived from these at a conclusion that testers have assumed a right "to define the potential for thinking, and then, regardless of the evidence, to draw broad conclusions about the extent to which a person possesses it" is a mystery.

Nairn questions whether the claims made for the tests "constitute an even minimally prudent construction of the facts." Certain of the evidence behind the statements that are made will be discussed below. But, it must be noted that Nairn's charge of misrepresentation is itself based on a calculated distortion of the record.

Use of Tests in Admissions

Consistent with its charges of misrepresentation, the Nairn report implies that scores on admissions tests are given undue weight, and that ETS is somehow an arbiter of admissions or "gatekeeper" to higher education. In fact, colleges and graduate and professional schools make their own admissions decisions, each institution using its own criteria. The criteria are decided upon by faculty and administrators at these institutions and in the case of many public institutions by state boards of education as well. The notion that ETS decides admissions is clearly a fallacy.

As indicated above, ETS and the sponsors of the admissions tests encourage institutions to weigh various kinds of information in admissions decision-making. But what are the facts about test use? Despite the advice institutions receive, do they generally place an undue weight on test scores?

At the college level, this does not seem generally to be the case. Most colleges are not selective and admit a large proportion of their

applicants. A very large majority of students intending to go to college do so. A 1972 study conducted for the federal government found that of high school graduates who applied to college, 87.5% had been admitted to at least one institution by the end of their senior year (Hilton and Rhett, 1973). A 1978 survey conducted for the American Council on Education indicated that 75% of freshmen were attending their first choice college and nearly 95% were attending their first or second choice college (Astin, King, and Richardson, 1978, p. 18).

Even among selective colleges test scores are seldom the most important factor in admissions, let alone the sole factor. A 1979 survey conducted jointly by the College Board and the American Association of Collegiate Registrars and Admissions Officers found that *under 2% of the selective colleges responding to the survey indicated that test scores were "the most important factor" in admissions decision-making.* Ninety percent described test scores as a "very important factor" or as "one of several factors" (Van Dusen, Nelson, Jacobsen, and Ivens, 1979, p. 26). These general survey results are no cause for complacency about the issue of proper test use. But, they do suggest that test scores are much less crucial in admissions decisions than Nairn supposes.

Much attention is given in the Nairn report to law schools where admissions are generally more competitive than at the college level. As noted in the Nairn report, the Association of American Law Schools stated in 1973: "for the first time in the history of United States legal education every accredited law school denied admission to applicants who it considered qualified for the study of law." Undoubtedly, many law schools could fill their classes several times over with qualified candidates selected from their applicant pools. This situation is a function of the number of places in law schools in relation to the number of individuals who want to study law, and it would continue to be true if admissions tests were abolished tomorrow. If the LSAT were removed from the scene, law schools would still be selective, but they would have less useful information on which to base selection decisions.

The Nairn report provides little evidence that LSAT scores are given too much weight in admissions. He suggests that a minority of law schools may use cutoff scores on the LSAT, but there is no indication whether these cutoff scores (if used) were related to minimum admissability requirements or were used to disqualify large numbers of applicants. Nairn discusses at length the fact that predictive indices combining test scores and previous grades are reported to law schools, but he fails to point out that these reports are accompanied by full copies of each student's academic transcript.

An overall picture is provided by a table on page 460 of the Nairn

report which he adapted from Willingham and Breland (1977). It indicates that admission to ABA accredited law schools is clearly related to students' undergraduate grades and test scores, but these factors are far from decisive. Only for students with the lowest or with the very highest grades and test scores are the probabilities of admission to at least one law school close to zero or one. The admissions policies of individual ABA-approved law schools are described in the *Prelaw Handbook* (1979) published by the Association of American Law Schools and the Law School Admission Council. These descriptions indicate that the vast majority of law schools consider such factors as recommendations, program of study, extracurricular activities and community service, work experience, personal qualities, accomplishments, and special talents as well as grades and test scores in making admissions decisions. However, a few public institutions give little weight to subjective assessments and appear to rely very heavily on previous academic grades and test scores.

ETS assists institutions in making appropriate use of test scores and other information through regular publications, advisory services, research, and other means. An example of the last of these is a paper prepared by ETS Senior Vice President Winton H. Manning (1977) for the Carnegie Council on Policy Studies in Higher Education. Manning proposed an alternate admissions model—one that would assist institutions in moderating the weight given to test scores. Under this structure, institutions would make decisions in two stages: first, identifying those students who meet minimum admissability requirements in terms of curriculum preparation, grades, and test scores; and second, selecting from among the admissable students those who in the light of all the academic and nonacademic evidence, would seem to best advance the educational philosophy and objectives of the institution, the profession, and society. There is no universal agreement on this model, though variants of it are undoubtedly in use in many institutions. But, it does represent a constructive way to help insure that multiple relevant values are considered in admissions.

What proposals does Nairn advance to help those institutions with many more applicants than places to make admissions decisions in a fair and wise manner? It is difficult to find in the Nairn report constructive and practical alternatives to the current use of tests as one element in admissions. The other kinds of information Nairn proposes for use, such as essays or personal statements submitted as part of the admissions process or reports of special talents and accomplishments, are already used by many selective institutions. However, these institutions also need uniform and dependable measures of students' developed abilities as an aid in judging which candidates have the best academic preparation and as one means of

comparing applicants in an equitable way. Nairn fails to consider the consequences that both institutions and applicants would suffer if admissions tests were not used.

In advancing his case against the LSAT specifically, Nairn overlooks the fact that the best available single predictor of academic performance in most law schools is the LSAT—that LSAT scores are very generally found to be more closely related to average law school grades than are previous college grades or any other known predictor. This finding does not suggest that LSAT scores should be used alone to decide admissions, for that is certainly inadvisable. But, is it not also inadvisable to disregard clearly relevant test-based information, and to base decisions entirely on information that research shows to be less predictive of future academic performance? There is no more logical basis for weighting LSAT scores 0% than there is for weighting them 100%.

Admissions, like most important decision problems, is a matter of values, judgment, and evidence. The admissions processes of selective educational institutions have been worked out through years of experience, consideration by responsible educators, debate, and research. No doubt, changes and improvements in admissions can be made. Nairn wants institutions to devalue information provided by tests. But, he does not suggest alternatives that would perform the important functions now performed by admissions tests. Because the Nairn report misrepresents the tests and their uses and because it offers no serious alternatives, it is unlikely to contribute constructively to improvement of admissions to higher education.

Evidence Supporting Test Use

Several kinds of evidence supporting use of the tests exist. At the foundation is an informed judgment by the designers and developers of the tests that the tasks sampled by the tests require skills that are important to competent academic performance—that the tests are in essence academic work samples. This judgment is guided by experience with particular types of questions or problems in other settings and by research showing a relationship between performance on these tasks and academic success. In addition, this judgment is subject to review of independent educators, particularly those in the institutions that use the tests. The practice of publishing sample test forms with answer keys for each of the major admissions testing programs further permits the widest possible scrutiny of test content. There is considerable agreement among those using the tests that the abilities assessed by the reading, verbal reasoning, and mathematical problems contained in the tests are relevant to successful academic work in their institutions. The prevailing view of the SAT is the one expressed by William Ambler, dean of admissions

at Haverford College (as quoted in *Newsweek*, February 18, 1980): "The test reflects the words and symbols that students must deal with in courses every day."

Nor is this judgment of content relevance a static one. Each of the admissions tests has changed over time. In each case, the new material has been thoroughly reviewed by representatives of institutions using the tests and has been subject to research on the empirical relationship of performance on the new types of tasks to success in academic work. (See, e.g., French, 1957; French, 1964; Flaugher and Rock, 1966; Schrader, 1973; McPeek, Pitcher, and Carlson, 1974).

Many technical studies of the tests are performed, including studies of test difficulty, the consistency of meaning of scores from one test form to another, the reliability (or stability and precision) of scores, and the relationships of scores on the tests to other variables. Studies of this last type are often concerned with the "validity" of a test—the extent to which scores relate to other measures of educational preparation, development, or achievement. Important among them are studies of predictive validity, or the extent to which test scores are related to future academic performance (usually measured by grades). It is from the results of these studies that Nairn attempts to show that use of the tests is little better than rolling dice.

Predictive validity of the admissions tests. Though evidence from predictive validity studies is by no means the whole story, it is important as an indication that test scores bear a meaningful relationship to future academic performance. Such studies cannot be conclusive, however, for in many cases predictive studies fail to reflect the full value of the tests to selective institutions. These institutions use test scores and other academic information to identify students who are likely to do well. To a large extent, the institutions are successful, and the great majority of admitted students are fully capable of meeting the requirements of these institutions' academic programs; students lacking the basic qualifications are seldom admitted. Since the prediction studies examine the relationships of scores to grades for a restricted, capable group of enrolled and persisting students, the observed relationships of tests and grades in these studies are lower than those that would be found for the total group.

Typical results of prediction studies based on the test scores and grades of enrolled students are shown in Table 1.

Table 1

Characteristic Validity Coefficients of Admissions Test Scores and Previous Grade Record (GPA) for Predicting Subsequent Grades

			Median Validity Coefficients		
Admission Test	Type of School	No. of Studies	Test Scores	Previous GPA	Both Predictors Combined
SAT	Undergraduate	827	.41	.52	.58
GRE	Graduate Arts & Sciences	24-30	.33	.31	.45
LSAT	Law	116	.36	.25	.45
GMAT	Graduate Management	57	.29	.21	.38

The numbers in the table are "correlation coefficients." A correlation coefficient is an index of relationship generally symbolized by the letter "r." An r of .00 indicates no relationship, (or more precisely no "linear" relationship). An r of 1.00 indicates a perfect relationship. It is standard professional practice to report predictive validity in terms of correlation coefficients.

The numbers shown in the table are median values; actual values vary from institution to institution. *For a variety of reasons (including the "restriction of range" problem described above), the numbers reported in Table 1 are conservative estimates.* These factors and their effects on validity coefficients are discussed later in this paper. In the discussion immediately following, however, data from the table will be used.

The numbers in Table 1 are generally the ones on which Nairn bases his claim that rolling dice is just about as good as using test scores in admissions. In fact, he uses a different and incorrect value for the characteristic validity of the SAT (.345) since he mistakenly averages the separate validities of the two parts of the SAT, rather than considering the validity of the total test.* Moreover, he decides to square the validity coefficients and then to multiply the results by 100 arriving at the numbers 12 for the SAT (based on the erroneous validity of .345), 11 for the GRE, 13 for the LSAT, and 8 for the GMAT. These he interprets as the percent of cases for which an ETS test will provide a prediction more accurate than a random prediction such as by throwing a pair of dice. Taking the complements of these numbers, he states that rolling dice will be as accurate as using test scores from 87% to 92% of the time. For the SAT specifically, Nairn

*The characteristic validity coefficients for the two parts of the SAT, verbal and mathematical, are .37 and .32, respectively. It does not require advanced statistical knowledge to recognize that the predictive effectiveness of the total test will be as high or higher than that for either part alone. If this were not so, one could simply use the part with the higher validity (e.g., the verbal test with a characteristic validity of .37) and obtain a higher validity coefficient than the average of the coefficients for the two parts (.345). But, in fact combining the information from the two parts results in greater predictive validity (typically .41) than does using either part separately.

says that "on the average, for 88 percent of the applicants (though it is impossible to know which ones) an SAT score will predict their grade rank no more accurately than a pair of dice." (p. 65)

It is impossible to know how Nairn arrived at this misconception, which was featured prominently both in the report itself and the press release that accompanied the report. The square of a correlation coefficient (or r^2) is a statistic often described as the "percent of shared variance" or occasionally as the "coefficient of determination." Little can be said of Nairn's interpretation except that it is wrong. It is safe to say that no reputable statistics text can be found anywhere that suggests that r^2 indicates the proportion of predictions better than chance predictions.

In fact, Nairn's interpretation makes no sense even at an intuitive level. If the SAT were invalid, it would be *no worse* than a random predictor, such as a pair of dice. If predictions based on an invalid test and random predictions were compared for a large group of students, the test would give predictions closer to students' obtained GPAs for about 50% of the cases, and random predictions would be better for the other 50%. Predictions based on valid information will be better than random predictions and will certainly more closely approximate obtained GPAs for more than 50% of the applicants. Yet, according to Nairn, dice will be as good as or better than the SAT, a test of proven validity, 88% of the time. Like citations of evidence in the Nairn report, these dice should be carefully inspected.

In what percent of the cases is it better to use an admission test instead of dice? A reasonable answer is 100%, for use of test scores in each case improves the *likelihood* of making an accurate forecast of future performance. Since predictions are inevitably somewhat uncertain, sometimes (but well under half the time) a blind guess will turn out by chance to have been more "accurate" than an estimate based on relevant information. But, this is hardly an argument for disregarding useful predictive information.

Despite Nairn's misinterpretation of the coefficient of determination (r^2), it is quite a respectable statistic which has a definite meaning to statisticians. However, there is agreement among many who have studied the matter that the coefficient of determination is not an appropriate measure of the usefulness of tests for selection. For example, in a widely read text, Cronbach and Gleser (1965, p.52) state:

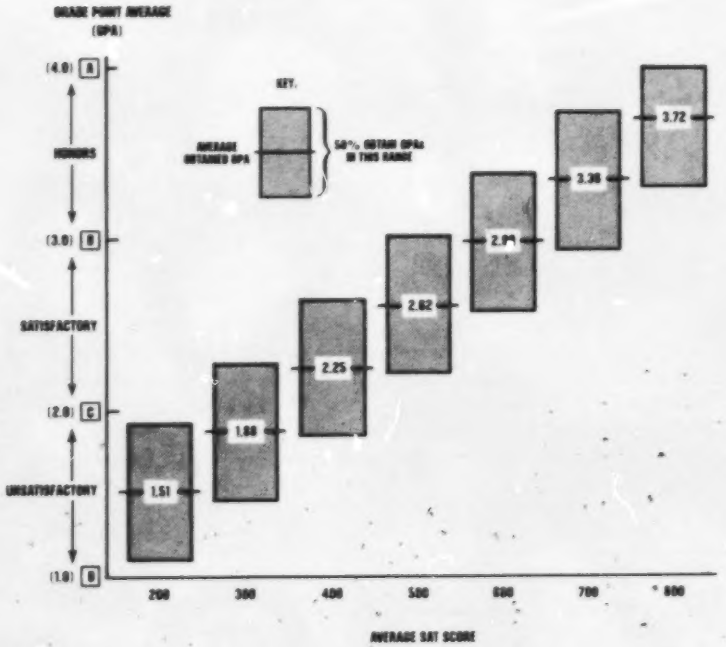
> Like others who have investigated the selection problem, we find no basis for interpreting the usefulness of tests in terms of the coefficient of forecasting efficiency or the coefficient of determination. These indices should not be employed in evaluating tests for selection purposes.

What can predictive validities tell us about the usefulness of tests for selection purposes? A college using the SAT may be particularly interested in the average grades that selected students will attain. Figure 1 illustrates the average grades that would be expected at a college typical of those conducting studies through the College Board's Validity Study Service in 1974. It is a composite illustration based on the typical distributions of test scores and grades found at these colleges and on a correlation of .40, which is just below the characteristic validity of the SAT. (In technical terms, it depicts the "linear regression" of grades on test scores.) While there is no way to turn a correlation coefficient directly into an easily interpreted percentage, Figure 1 indicates that the relationship of SAT scores to grades is hardly random—those students with higher SAT scores are typically those who go on to earn higher grades in college. As shown in the figure, there is also variation in the grades obtained by

FIGURE 1

AVERAGE COLLEGE GRADES
FOR STUDENTS WITH DIFFERENT SAT SCORES

TYPICAL SAT SCORE-COLLEGE GPA CORRELATION = .40
BASED ON DATA FROM 150 VALIDITY STUDIES IN 1974



18

students with any given score. The figure shows grades for the middle 50% of the students at each score level. Other students will obtain scores outside these ranges. While predictions are not certain, college faculties and admissions officers find the relationship of test scores to grades depicted in Figure 1 agrees with their own experience, and they find using SAT scores with grades and other information in admissions much preferable to rolling dice.

The usefulness of a test can also be evaluated in "percentage correct" terms by determining the proportion of selected or unselected students who would perform above a given criterion level. For example, consider a college with 1000 applicants, but only 900 places. It has to deny admission to 100 students. If these students were chosen at random, about 50 of the students denied admission would have ranked in the top half of the class in terms of later grades. Use of a test with a validity of .41 to select these students would reduce the number of incorrect decisions to 21. Use of test scores and previous grades with a combined validity of .58 would reduce the number to 11. These results are based on a standard set of statistical tables developed by Taylor and Russell (1939). This approach to evaluating validity in terms of the likely number or percentage of incorrect decisions is known to Nairn, who refers to it in a footnote on pages 419-420.

The tests' contribution to prediction. In considering what tests add to prediction, Nairn concludes that "inclusion of SAT scores in the prediction process improves the prediction of college grades by an average of only five per cent or less" (p. 66). For example, in a table on page 67 he shows an "improvement" of 5.1 based on data from 1974 validity studies. Simplifying matters a bit, he states: "This thin margin . . . is the single thread—the single rational function—on which the ETS aptitude empire hangs" (p. 66).

In arriving at a figure of five percent, which in his chapter title he calls "five percent of nothing," Nairn switches statistics—from r^2 to another index. In fact, he uses the other statistic referred to by Cronbach and Gleser, above, the "coefficient of forecasting efficiency," which is equal to $1-\sqrt{1-r^2}$. The actual coefficients for the 1974 data (not reported by Nairn) are 13.4 for high school grades alone and 18.5 for tests and grades combined, which do indeed differ by 5.1. This index has a theoretical range from 0 (for random prediction) to 100 (for perfect prediction) and can be interpreted in percentage terms, though the interpretation is not a particularly simple one. Since "perfect prediction" cannot be attained, it is reasonable to consider the contribution of tests in relation to the level of prediction offered by grades alone. Simple arithmetic (5.1 divided by 13.4) indicates that combining test scores with grades increases the value of Nairn's chosen index *by 38%, not 5%*.

Nairn does not attempt a similar demonstration for other tests,

such as the LSAT, because he says (without presenting any data) that "the relative predictive ability of grades and scores swings back and forth from one time period to another" (p. 66). If he had simply taken the characteristic law school validities and had computed his index values from them, he would have found coefficients of 3.2 for previous grades alone and 10.7 for grades and LSAT scores combined—a difference of 7.5 and a percentage improvement of 234%. It is hard to understand why Nairn does not do this, because he uses the LSAT validities from Table 1 of this paper (in modified form) in other sections of his report.

In Nairn's original terms of reference (r^2) it is also difficult to arrive at a 5% improvement figure. Using the correlations from Table 1 of this paper, the results are approximately as follows for the four admissions tests:

	r^2 for previous GPA	r^2 for tests and GPA combined	Increase in r^2	Percentage improvement
SAT	$.52^2 = .27$	$.58^2 = .34$	.07	$.07/_{.27} = 26\%$
GRE	$.31^2 = .10$	$.45^2 = .20$	.10	$.10/_{.10} = 100\%$
LSAT	$.25^2 = .06$	$.45^2 = .20$	.14	$.14/_{.06} = 233\%$
GMAT	$.21^2 = .04$	$.38^2 = .14$	.10	$.10/_{.04} = 250\%$

It should be clearly understood that ETS makes no claim that the LSAT improves prediction by 234% or the GMAT by 250%, though interpretations in terms of Nairn's varying statistical indices lead to these conclusions. These results, in fact, are not helpful in understanding the relative contribution of grades and tests to prediction.

It is important to recognize that test scores and previous academic records each contribute uniquely to prediction. One meaningful way to assess the contribution of several variables to prediction is to examine the weights that must be applied to each variable to obtain optimum overall prediction. Consider a college at which SAT scores alone have their typical validity of .41 and where high school grades alone have their typical validity of .52. In using these predictors together to obtain a typical overall predictive validity of .58, the test scores would be given 39% of the weight and grades would be given 61% of the weight. At the graduate and professional school level, the validity study results would lead one to place even more weight on tests than on grades in order to obtain optimum prediction. But, this is not a decision that should be based solely on the results of validity studies.

Prior grades reflect the pooled judgment of a number of instructors who have evaluated students' work on tests, term papers, class projects, recitations, and other aspects of academic work. Admissions tests reflect student performance on a carefully selected sample of intellectual tasks, scored in a highly consistent way across individuals. Students' prior academic records are properly given great weight in admissions. Likewise, scores on the tests are properly weighted in these decisions. Although tests measure a narrower range of abilities than grades, they are also valid. Because test scores have a consistent meaning for all students, they help to redress the unfairness that would occur if grades from a variety of sending schools with different grading standards were used alone. It is very largely this fairness, as well as a desire to select students who will perform well academically, that motivates institutions to use the tests.

What is obscured in the Nairn analysis is that deciding on admissions criteria is fundamentally a matter of judgment. These judgments take into account the relevance, dependability, comparability across individuals, comprehensiveness, and validity of the information considered for use. It is not a matter that should be decided in simple, absolute terms relying entirely on the results of limited statistical studies. If the decision were made in these terms, test scores would probably be given greater weight than is typically the case currently, for the studies show them to be nearly as valid as or, in many cases, more valid than grades. Fortunately, admissions officers and admissions committees generally take a restrained view of validity study results and use considerable judgment and discretion in weighing other kinds of evidence.

Problems in interpretation. As noted earlier, the estimates of predictive validity provided by the validity studies are conservative estimates. This is due, in part, to the effects of selection and the resulting restriction in range of abilities for selected groups. Formulas for adjusting for restriction of range are available, but their use is based on untested assumptions. Practical examples show, however, that the effects of restriction in range can be severe. A dramatic example can be found in a report by Thorndike (1947, p. 66) on World War II selection of Army Air Force pilots. The entire group of would-be pilots took the test, and the entire group was allowed to proceed through training before a final selection was made. Validity of the test was .18 for the selected group; for the unselected group as a whole, the validity was .64.

An example closer to the college admissions scene can be found in a study by Jackson (1977) which reported the correlations of SAT scores with *high school* grades. For groups of students selected by and enrolled in colleges participating in College Board validity studies, the typical correlations of high school grades with SAT verbal

and mathematical scores were found to be .29 and .31, respectively. For general samples of SAT test takers, the typical correlations between self-reported high school grades and SAT verbal and mathematical scores were .50 and .53, respectively. (Note that these correlations are for separate scores on the two parts of the SAT. The typical correlation of combined SAT scores with high school grades for unselected groups of high school juniors and seniors is .56.)

Another factor that attenuates the validity coefficients is "criterion unreliability"—the fact that the grades which are predicted are imperfect measures of academic performance. Students take different courses and programs of study. Grading standards vary from program to program and from instructor to instructor, and even the grade a student receives from a particular instructor in a particular course is generally a matter of judgment. The grade point averages that students attain are used as criterion measures in validity studies because they are useful and reasonable outcome measures and because they are meaningful to faculty members, not because they are viewed as ideal or perfectly consistent measures of academic performance in college.

Again, formulas are available for adjusting validity coefficients for criterion unreliability, but they are seldom used because the degree of criterion unreliability is not precisely known. However, empirical studies show that the effects on validities may be substantial. At the University of California, Goldman and Slaughter (1976) estimated what the combined validity of high school grades and SAT scores would be if GPAs for all students were based only on grades earned in the same college courses. They found a marked increase in validity (from about .44 to .70). Because of technical difficulties in this study and because of uncertainty about the degree to which results can be generalized, the results may overstate the typical effects of "criterion unreliability" on predictive validity. Nevertheless, these results and those of previous research by the senior author led these researchers to conclude: "In sum we believe that the validity problem in GPA prediction is the result of the GPA criterion, rather than the tests used as predictors."

The two factors discussed here—restriction of range and criterion unreliability—probably explain, in large part, why performance at the graduate and professional level appears to be less predictable than performance at the college level. The range of talent at the graduate level is more restricted than at the college level, and grades are generally concentrated into only a few categories (e.g., A, B and C), instead of the five categories that are more often used at the undergraduate level.

Nairn seems to be well aware of these problems in interpretation of validity study results and of the standard approach to research on test validity, for there is a discussion of these matters in a footnote

on pages 416-420. Unaccountably, he dismisses these concerns without stating clearly why they are unworthy of consideration. He is correct, however, in stating that testing agencies avoid adjusting computed validity coefficients, since it is judged preferable to report the more conservative estimates which are understood as such by professional users.

Standards for validity. What are reasonable standards for test validities and to what extent do the major admissions tests meet these standards? Because ideal validity studies cannot be performed, the validities are not precisely known. Even without taking into account the various factors that attenuate validity coefficients, however, the obtained results indicate that the tests are useful predictors. Figure 1 illustrates that a test with a validity of .40 is quite useful in identifying students who are *likely* to do well in academic course work at the college level.

The predictive validities of tests can also be judged in relation to those of the usual alternative predictors. At the college level, the characteristic validity of the SAT (.41) is not a great deal lower than that of high school grades (.52). In fact, these values vary from college to college, and in about 25% of the colleges the SAT is a better predictor than high school grades. At the graduate and professional levels, as noted, the tests most often have somewhat higher validities than does undergraduate GPA. *On the whole, the tests are about as useful in predicting future academic performance as previous academic grades, the predictor that is most widely used and accepted.*

Nairn prefers to judge the test validities in relation to "perfect prediction" which is, of course, unattainable. Human behavior and performance are notoriously variable, many factors (academic and personal) affect students' achievement, and grades are an imperfect measure of academic outcomes. Anything approaching perfect prediction, in fact, would require massive intrusion into people's lives to obtain the needed predictive data and strict regimentation of later academic experiences and other aspects of life. Colleges do not seek certainty in their admissions programs. Instead, they attempt to make the best judgments possible using the most relevant available prior information. Nairn believes this is not good enough, but he does not indicate what educational institutions, faced with a need to make intelligent decisions, should do.

It may be recalled that Nairn expressed doubt about whether the claims made for the tests are supportable. If it were claimed that the tests are perfect or if it were recommended that test scores be used alone to decide admissions, there would be grounds for grave concern. But, the claims made in test program publications are that the tests measure particular skills in the areas of reading, vocabulary, verbal reasoning, and mathematics that have been

23

shown to be related to future academic performance, it is recommended that test scores be used to supplement students' prior academic records. In view of the judgment by educators that the tests do measure important academic skills and in view of abundant research which clearly shows a useful predictive relationship of test scores to later grades, it is very difficult to conclude that Nairn's charges of fraud and misrepresentation themselves "constitute an even minimally prudent construction of the facts."

Beyond Academic Selection

Though Nairn gives much attention to the issue of test validity, he seems to recognize that his demonstration is unconvincing, for he also criticizes college grades, the criterion predicted by the SAT (pages 80-82). In his catch phrase—"5% of nothing"—the word "nothing" represents college GPA. It is not a matter of imperfect measurement but what GPA stands for. Here again there seems to be a difference between Nairn's values and those of the colleges and universities, a difference far more fundamental than any dispute over predictive validity. Colleges clearly value academic talent and excellence as qualities to be sought and nurtured. Nairn's values are not precisely clear, but his report gives little attention to the importance of learning and academic achievement. This may help to explain why Nairn's views on how to go about admissions are so divergent from those of college and university faculties.

These comments are perhaps somewhat unfair to Nairn for probably he does share many of the values that his mentor, Ralph Nader, expresses in a preface to the report. There Nader indicates that he would prefer to give more attention to characteristics not measured by multiple-choice tests, such as "judgment, wisdom, experience, creativity, idealism, determination, or stamina" (p. xii). Elsewhere in the report, Nairn suggests that more weight should be given to non-academic activities and accomplishments.

Though few in the colleges would agree with abandonment of current academic measures, there is unquestionably much sentiment that other personal qualities like those cited by Nader should be given more systematic attention in admissions. Though these qualities are highly valued and sought, they are difficult to assess in any but the most subjective terms.

ETS and the client groups it serves have for many years funded, conducted, and published research designed to help colleges make better use of information on personal qualities in admissions. (See, for example, Anastasi; Meade, and Schneiders, 1960; Davis, 1964; and Baird, 1979, several of which are cited in the Nairn report.) Currently the College Board and ETS are jointly engaged in a major research project in collaboration with nine colleges to develop

improved assessments of personal qualities, special talents, skills, and significant accomplishments. As part of this project, the relationships of these measures to student development over the four college years—development broadly conceived in both academic and non-academic terms—will be studied.

ETS has no quarrel with Nader and Nairn's advancing their views concerning appropriate criteria for use in selective admissions. But, it cannot be agreed that Nader and Nairn's values are the only values, or that these values should be substituted for the considered judgment of the faculties and administrators of the educational institutions concerning the proper weighing of academic and non-academic information in admissions. Nor are the tactics employed by Nader and Nairn in advancing their point of view designed to lead to thoughtful debate of these issues.

Institutions using admissions tests believe that these tests have important benefits both to themselves and to their applicants. They do not view this as a narrow technical issue, though in large part statistical research confirms that the tests are useful. They are also well aware that the information and assessments of student characteristics they use in admissions are not infallible. Opportunities to contribute constructively to improvement of testing, assessment and admissions are great. The Nairn report, which misrepresents many important facts and which offers no serious alternatives, contributes very little to such improvement.

References

Anastasi, A., Meade, M. J., & Schneiders, A. A. *The validation of a biographical inventory as a predictor of college success.* College Entrance Examination Board Research Monograph No. 1. New York: College Entrance Examination Board, 1960.

Angoff, W. H. (Ed.). *The College Board admissions testing program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests.* New York: College Entrance Examination Board, 1971.

Astin, A. W., King, M. R., & Richardson, G. T. *The American freshman: National norms for fall 1978.* Los Angeles: University of California, Los Angeles and the American Council on Education, 1978.

ATP guide for high schools and colleges, 1979-81. New York: College Entrance Examination Board, 1979.

Baird, L. L. *Development of an inventory of documented accomplishments for graduate admissions.* GRE Board Research Report GREB No. 77-3R. Princeton, NJ: Educational Testing Service, 1979.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions.* Urbana: University of Illinois Press, 1965.

Davis, J. A. *Faculty perceptions of students. I. The development of the student rating form.* Research Bulletin 64-10. Princeton, NJ: Educational Testing Service, 1964.

Flaugher, R. L., & Rock, D. A. *The wide range validity of certain new aptitude tests.* College Entrance Examination Board Research and Development Report 66-7, No. 8. Princeton, NJ: Educational Testing Service, 1966.

French, J. W. *Validation of the SAT and new item types against four-year academic criteria.* Research Bulletin 57-4. Princeton, NJ: Educational Testing Service, 1957.

French, J. W. New tests for predicting the performance of college students with high-level aptitude. *Journal of Educational Psychology*, 1964, 55, pp. 185-194.

Goldman, R. D., & Slaughter, R. E. Why college grade point average is difficult to predict. *Journal of Educational Psychology*, 1976, 68(1), 9-14.

Hilton, T. L., & Rhett, H. *Final report: The base-year survey of the national longitudinal study of the high school class of 1972. Appendix B, Part II.* Washington, DC: National Center for Educational Statistics, 1973.

Jackson, R. Correlations of SAT scores with high school record. Appendix, *On further examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline.* New York: College Entrance Examination Board, 1977.

McPeek, W., Pitcher, B., & Carlson, A. *The predictive effectiveness of several experimental item types and the operational item types in the Law School Admission Test in 1970-71.* Law School Admission Council Research Report 74-1. Princeton, NJ: Educational Testing Service, 1974.

Manning, W. H. The pursuit of fairness in admissions to higher education. In *Selective admissions in higher education* (a report of the Carnegie Council on Policy Studies in Higher Education). San Francisco: Jossey-Bass, 1977, pp. 20-64.

Nairn, A., & Associates. *The reign of ETS: The corporation that makes up minds.* Washington, DC, 1980.

Prelaw handbook: Annual official guide to ABA-approved law schools. Association of American Law Schools and the Law School Admission Council, 1979.

Schrader, W. *Validity of the quantitative comparison test.* Statistical Report, SR 73-60. Princeton, NJ: Educational Testing Service, 1973.

Taking the SAT. New York: College Entrance Examination Board, 1978.

Taylor, H. C., & Russell, J. T. The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, 1939, 23, 565-578.

Thorndike, R. L. (Ed.) *Research Problems and Techniques: Report No. 3.* Army Air Forces Aviation Psychology Program Research Reports. Washington, DC: U.S. Government Printing Office, 1947.

Van Dusen, W. D., Nelson, J. E., Jacobson, E. C., & Ivens, S. H. *The College Board—AACRAO survey of undergraduate admissions policies, practices, and procedures: A special report on admissions requirements and test use.* New York: College Entrance Examination Board, 1979.

Willingham, W. W., & Breland, H. M. The status of selective admissions. In *Selective admissions in higher education* (a report of the Carnegie Council on Policy Studies in Higher Education). San Francisco: Jossey-Bass, 1977, pp. 65-252.